Advances in Applied Mathematics 应用数学进展, 2024, 13(3), 891-899
Published Online March 2024 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2024.133084

Hans汉斯

基于改进坐标注意力和U-Net网络的高分辨率遥感图像建筑物提取

陈　康

成都理工大学数理学院，四川　成都

收稿日期：2024年2月13日；录用日期：2024年3月8日；发布日期：2024年3月13日

摘　要

在城市规划、统计调查和灾害应急评估等领域，从遥感图像中准确提取建筑物至关重要。然而，由于高分辨率遥感图像中建筑形态的多样性和地面环境的复杂性，实现建筑的完整、高精度提取仍然是一个挑战。为此，本文提出了一种用于从高分辨率遥感图像中提取建筑物的新网络，该网络保留了U-Net的编码器－解码器结构，并融合了坐标自注意模块(CSAM)，以调整网络对输入图像中不同区域的关注程度，使得网络能够有选择性地捕捉和强调重要的语义信息，增强特征提取能力。在空间分辨率为0.3 m的WHU建筑物数据集上进行的实验结果表明，与U-Net、PSPNet、DeepLabV3+相比，所提出的网络能够获得更准确的建筑提取结果，达到98.21%的像素精度、95.28%的精准率、94.57%的召回率和90.34%的交并比。

关键词

注意力机制，U-Net网络，语义分割，建筑物，高分辨率遥感图像

Building Extraction from High-Resolution Remote Sensing Images Based on Improved Coordinate Attention and U-Net Network

Kang Chen

College of Mathematics and Physics, Chengdu University of Technology, Chengdu Sichuan

Received: Feb. 13th, 2024; accepted: Mar. 8th, 2024; published: Mar. 13th, 2024

Abstract

Accurately extracting buildings from remote sensing images is crucial in areas such as urban

文章引用: 陈康. 基于改进坐标注意力和 U-Net 网络的高分辨率遥感图像建筑物提取[J]. 应用数学进展, 2024, 13(3): 891-899. DOI: 10.12677/aam.2024.133084

planning, statistical surveys, and disaster emergency assessment. However, due to the diversity of building forms and the complexity of ground environment in high-resolution remote sensing images, achieving complete and high-precision extraction of buildings remains a challenge. Therefore, this paper proposes a new network for extracting buildings from high-resolution remote sensing images, which retains the encoder decoder structure of U-Net and integrates a Coordinate Self Attention Module (CSAM) to adjust the network's attention to different regions in the input image, enabling the network to selectively capture and emphasize important semantic information and enhance feature extraction capabilities. The experimental results on the WHU building dataset with a spatial resolution of 0.3 m show that the proposed network can achieve more accurate building extraction results compared to U-Net, PSPNet, and DeepLabV3+, achieving pixel accuracy of 98.21%, accuracy of 95.28%, recall of 94.57%, and intersection to union ratio of 90.34%.

Keywords

Attention Mechanism, U-Net Network, Semantic Segmentation, Buildings, High-Resolution Remote Sensing Images

1. 引言

在城市规划、统计普查和灾害应急评估等领域，快速准确地从遥感图像中提取建筑物具有重要意义[1]。然而，随着遥感技术的进步，高分辨率遥感图像中包含了更详细的土地覆盖信息，建筑形态的多样性和地面背景的复杂性，影响了建筑提取的准确性和完整性。因此，利用自动化的图像分割方法来实现高精度、完整的建筑提取仍然是一项具有挑战性的任务。

21 世纪以来，图像分割作为一个研究热点受到了广泛关注，人们提出了许多分割方法，通常分为两类：基于空间和特征的传统方法，以及基于深度学习的语义分割方法。传统的图像分割方法通常是基于特征域和空间域，利用图像中分割目标的形状、纹理等先验信息来获得分割结果。Kalyankar 等人[2]使用五种不同的阈值分割算法对遥感卫星图像进行分割，并对其分割效果进行比较。Liow 等人[3]采用边缘检测定位建筑物边界，然后采用目标区域增长确定建筑物的位置和面积。Avudaiammal 等人[4]将形态、光谱和几何特征信息与支持向量机(SVM)相结合，利用形态建筑指数将遥感图像分为建筑物和非建筑物。然而，传统的图像分割方法通常需要手工设计特征，费时费力，且泛化能力有限。

近十年来，神经网络不断发展，图像分割领域开始将卷积神经网络(Convolutional Neural Network, CNN)作为主流研究方法。2015 年，Long 等人[5]提出了全卷积网络(Fully Convolutional Network, FCN)，这是第一个用于像素级预测的端到端的图像语义分割网络。FCN 的成功使得一系列基于 CNN 的语义分割网络如雨后春笋般涌现出来，如 SegNet [6]、U-Net [7]、PSPNet [8]和 DeepLab 系列[9] [10] [11] [12]。得益于 CNN 在遥感领域的广泛应用，许多基于 CNN 的方法被应用于提高建筑提取的准确性和效率。李传林等人[13]将 ResNet 中的残差模块和注意力机制与 U-Net 有效结合，使网络在充分利用特征信息的同时，有效消除随机噪声对建筑物提取精度的影响。Qiu 等人[14]提出的 Refine-Unet，通过加入 IDSC 和 ASPP 模块改善跳跃连接，增强多尺度的特征提取能力，提高了建筑物的提取精度。然而，由于感受野受限，传统的卷积神经网络无法有效捕获全局上下文信息，在建筑物提取上仍然存在一些问题。

在上述工作的基础上，受 U-Net 架构思想的启发，提出了一种改进网络。本文的主要工作如下：1) 提出了一种从高分辨率遥感图像中准确提取建筑物的新网络，该网络保留了 U-Net 的编码器－解码器结构，并加入了注意力机制以提高特征提取能力；2) 设计了一种坐标自注意模块(CSAM)，它不仅可以捕获跨通道信息，还可以捕获位置敏感信息，以应对通道注意力对位置信息的忽视和空间注意力对捕获远程依赖关系的不足；3) 基于 WHU 建筑物数据集，验证了所提出方法的有效性，并与其他基于深度学习的方法进行了比较。

2. 方法

2.1. 编码器－解码器结构

编码器－解码器结构是深度学习图像语义分割领域中一种常见的网络架构，其设计灵感来自于需要在提取图像语义信息的同时保留图像空间信息的要求。采用编码器－解码器结构的典型图像分割模型有 SegNet、U-Net 和 DeepLabv3+等等，这些模型通过这一结构实现对图像语义信息的有效提取和像素级别的语义分割。

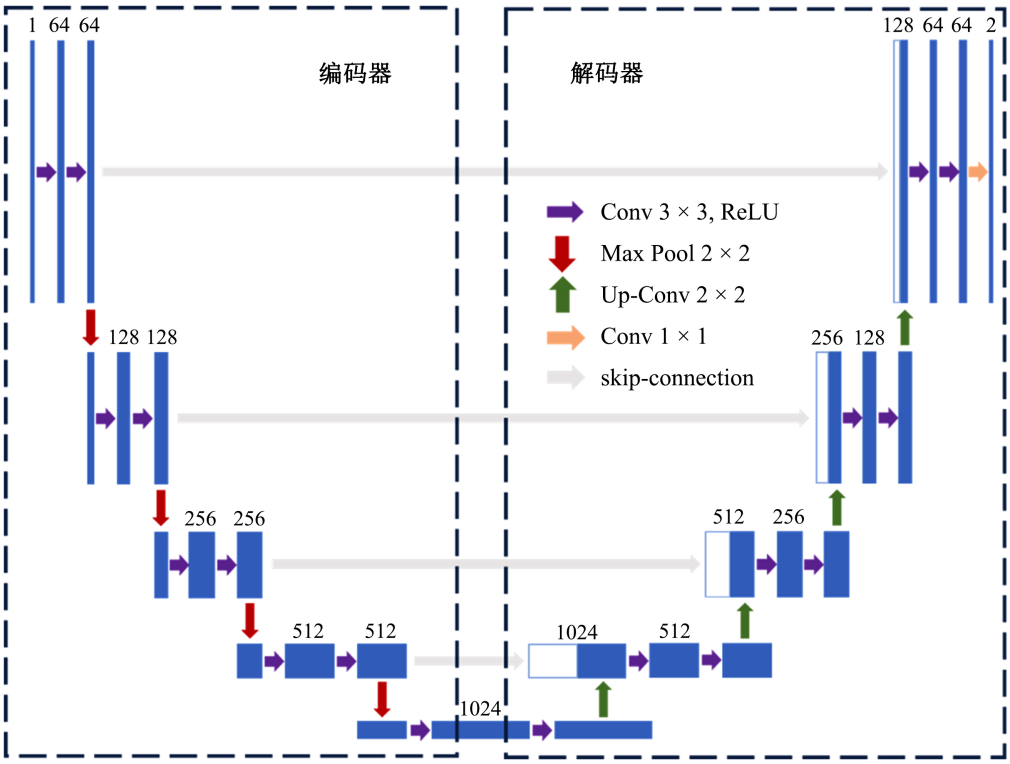


Figure 1. U-Net network structure
图 1. U-Net 网络结构

图 1 展示了 U-Net 的网络结构。左侧为编码器路径，负责从输入图像中提取层次化的特征表示，将图像信息转化为高级的语义表示，由 5 个单元组成，每个单元包含 2 个 3 × 3 卷积层，用于捕捉图像的局部特征，整个编码器路径含有 4 个 2 × 2 最大池化层，用于逐步降低特征图的空间分辨率；右侧为解码器路径，负责逐步恢复特征图的空间分辨率，将编码器提取的高级语义信息转化为像素级别的预测，生成与输入图像同尺寸的语义分割结果，同样由 5 个单元组成，每个单元同样包含 2 个 3 × 3 卷积层，用于进

一步提炼和细化特征，整个解码器路径含有 4 个 2 × 2 上采样层，用于逐步放大特征图，最后使用一个 1 × 1 卷积层将每个像素分为两类，输出最终的语义分割结果。U-Net 的独特之处在于引入了跳跃连接(skip-connection)，将编码器路径的特征图与对应解码器层的特征图进行拼接操作，实现特征融合，充分利用图像的浅层特征，减轻由于多次池化操作引起的信息损失，成功地解决了语义分割任务中的信息传递和细节保留问题。

2.2. 坐标自注意模块设计

在语义分割网络中，注意力机制被广泛应用以提高模型对重要语义信息的关注度，改进图像分割性能。SENet [15]是一个成功的例子，它通过对每个通道的特征进行挤压和激发来调整通道的注意力权重，提高模型对于重要通道的关注度。CBAM [16]将通道注意力和空间注意力相结合，在通道和空间维度同时对特征图进行加权，提高模型对图像不同位置和不同通道的关注程度，以更好地捕获语义信息。

受坐标注意力(Coordinate Attention, CA) [17]的启发，本文设计的坐标自注意模块如图 2 所示。首先将尺寸为 C × H × W 的输入特征图在宽和高两个方向上分别进行平均和最大池化，沿水平坐标和垂直坐标对每个通道进行编码，得到尺寸为 C × H × 1 和 C × 1 × W 的两组特征图；然后使用 1 × 1 卷积层、批量归一化(Batch Normalization, BN)和 ReLU 激活函数对分别聚合了高和宽两个方向特征信息的特征图进行通道压缩并进一步提取特征；接着再利用 1 × 1 卷积层和 Sigmoid 函数还原特征图的通道数并生成高和宽两个维度的注意力权重图，对应图 2 的 Part 1 和 Part 2。将两个维度的注意力权重相乘得到包含输入特征图坐标信息的尺寸为 C × H × W 的权重矩阵，最后将权重矩阵与输入特征图相乘得到坐标自注意模块的输出。坐标自注意模块可以在一个空间方向上捕获远程依赖关系，同时在另一个空间方向上保持精确的位置信息，使模型更准确地定位感兴趣对象的确切位置。

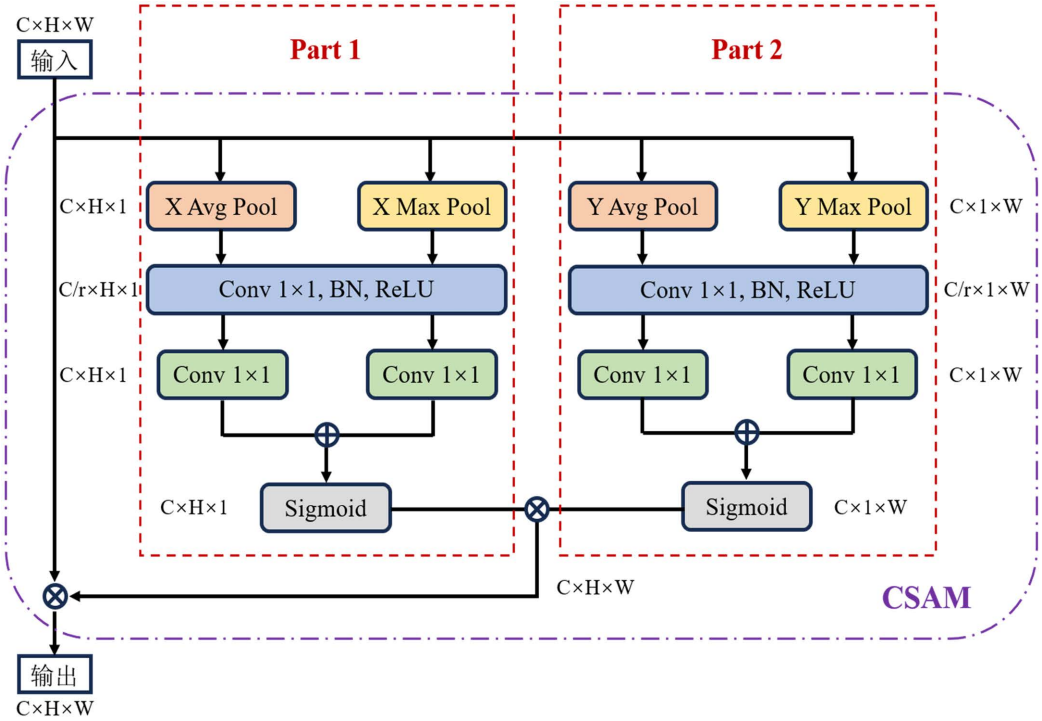


Figure 2. Coordinate self attention module
图 2. 坐标自注意模块

2.3. 总体网络架构

基于 U-Net 的编码器–解码器结构，本文提出的用于高分辨率遥感图像建筑物提取的改进 U-Net 模型结构如图 3 所示。将高分辨率遥感图像输入编码器路径，经过卷积和最大池化操作后，会得到 5 个不同层级的特征图，浅层卷积输出具有低级表层信息的高分辨率特征图，深层卷积输出的低分辨率特征图具有高级抽象信息。通过跳跃连接将编码器路径中的浅层特征与解码器路径中对应的特征图进行拼接，在特征融合阶段使用 CSAM，使网络通过学习的方式自适应获得坐标权重来调整模型对输入特征图不同位置的关注度，以增强有效信息，抑制无用信息，从而提高模型在建筑提取任务中的性能。每次特征融合后，再使用两次卷积操作进一步提取特征，最后使用一个 1 × 1 卷积层输出最终的建筑分割图。

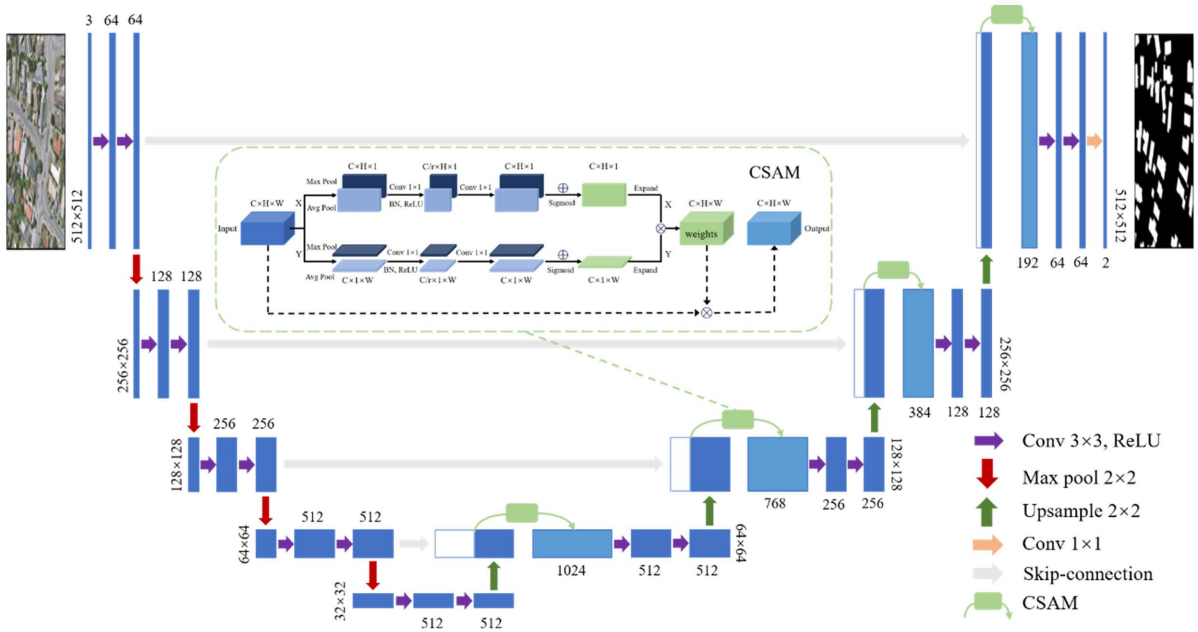


Figure 3. Model structure of this article
图 3. 本文模型结构

3. 实验结果与分析

3.1. 实验数据

实验在 WHU 建筑物数据集[18]上进行。如图 4(a)所示，WHU 建筑物数据集包含从新西兰克赖斯特彻奇城 450 平方公里的航拍图像中提取的 187,000 栋独立建筑，空间分辨率为 0.3 m。整张图像被无缝地裁剪成 8188 块，其中 4736 张为训练集，1036 张为验证集，2416 张为测试集，每张图片的尺寸为 512 × 512 像素。WHU 建筑物数据集由于空间分辨率高、标注精确等优点，被广泛应用于建筑物提取研究。

标签数据使用像素级的两个语义类提供，包括建筑物(白色)和非建筑物(黑色)，这两个语义类仅针对训练数据集公开。其中一组遥感图像及其对应的标签如图 4(b)、图 4(c)所示，它清晰地展示了具有规则和不规则形状屋顶的典型建筑。

3.2. 训练过程与评价指标

实验环境为 Windows 10 操作系统，内存为 32GB，使用 Nvidia GeForce RTX 4070 显卡加速计算，显存为 12 GB。本文所涉及模型均使用 Python 语言实现，使用的深度学习框架为 Pytorch，每个模型训练

100 个 epoch，每批次输入 8 张图像，损失函数采用交叉熵损失，使用 Adam 优化算法对损失函数进行优化，初始学习率设置为 0.0001，为了避免陷入局部最优，使用余弦退火策略进行学习率衰减。

本文使用像素精度(Pixel Accuracy, PA)、精准率(Precision)、召回率(Recall)和交并比(IoU) 4 个指标来评估模型性能。它们的计算公式如式(1)、(2)、(3)、(4)所示：

$$\mathrm{PA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{1}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{2}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{4}$$

其中，与真实标签相比，真正类(TP)、假正类(FP)、真负类(TN)和假负类(FN)分别表示正确提取建筑物的数量、错误提取建筑物的数量、正确提取非建筑物的数量和错误提取非建筑物的数量。

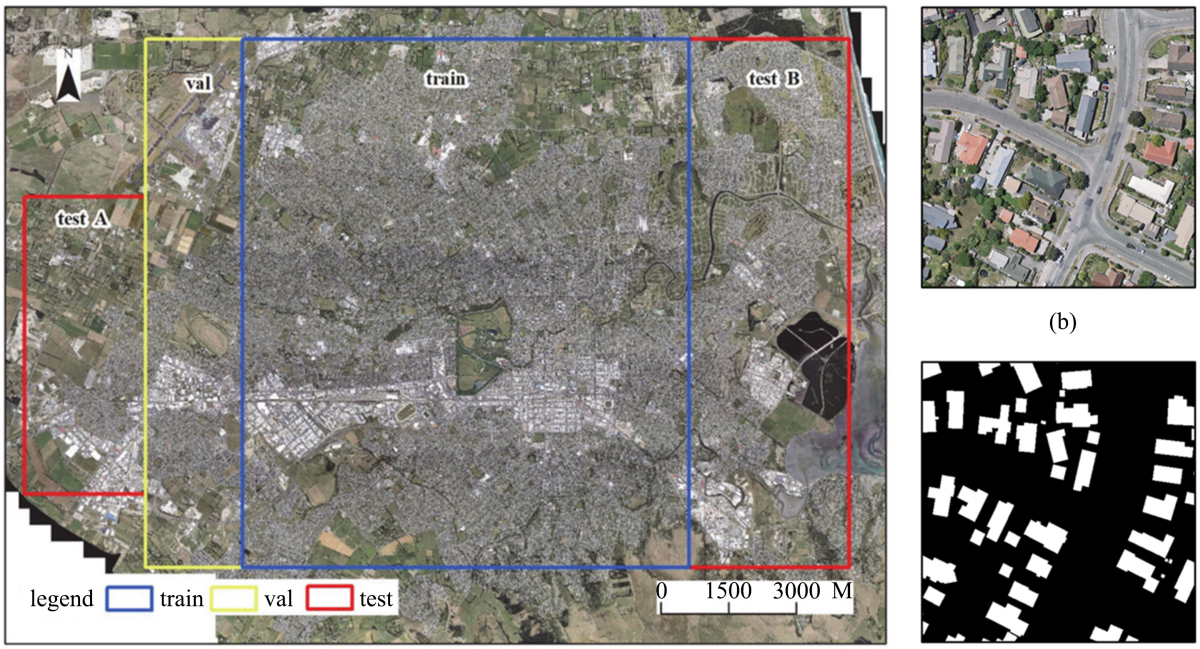


Figure 4. WHU building dataset. (a) Research area, the blue box represents the training area, the yellow box represents the validation area, and the red box represents the testing area; (b) Size 512 × 512 pixel original image; (c) The label of the original image

图 4. WHU 建筑物数据集。(a) 研究区域，蓝色框为训练区域，黄色框为验证区域，红色框为测试区域；(b) 尺寸为 512 × 512 像素的原始图像；(c) 原始图像的标签

3.3. 结果与分析

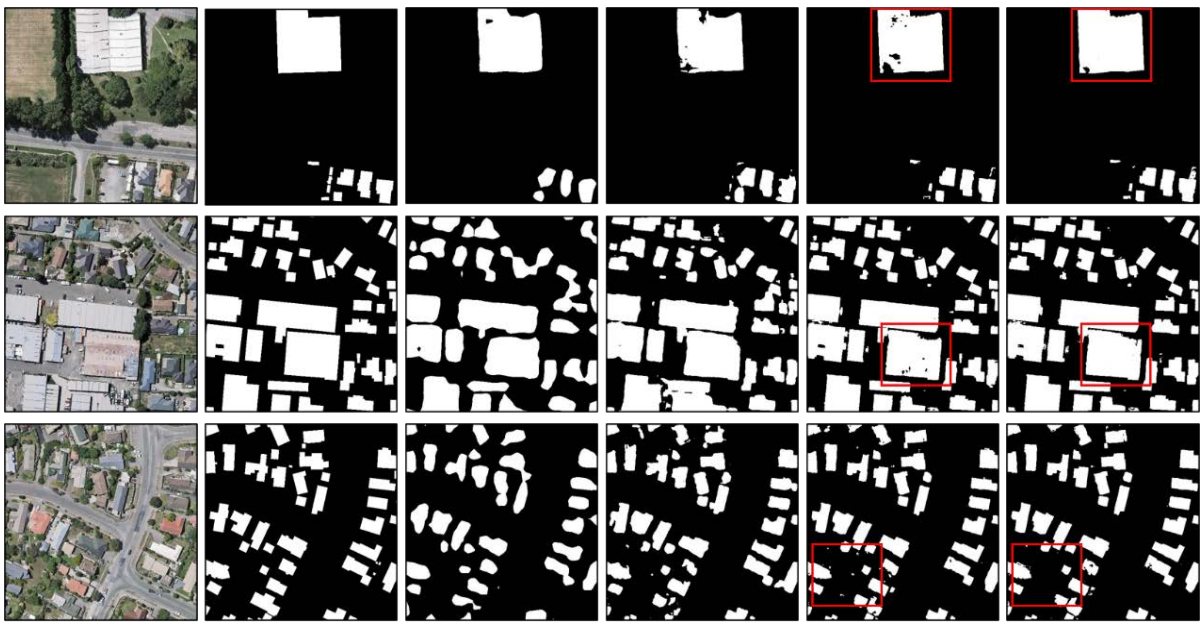
为了进一步验证本文模型在建筑物提取中的可行性和有效性，我们选取了 PSPNet、DeepLabv3+和 U-Net 等几种先进的语义分割模型进行比较，定量分析结果见表 1。在比较的方法中，U-Net 的指标结果要明显优于 PSPNet 和 DeepLabv3+，建筑物提取的像素精度、精准率、召回率和交并比分别到达了 97.76%、

94.11%、93.13%和 88.00%，在加入了 CBAM 和 SE 注意力模块后，U-Net 的建筑物提取精度有小幅提升。相比于 U-Net，本文模型在像素精度、精准率、召回率和交并比上分别提高了 0.45%、1.17%、1.44%和 2.34%。

Table 1. Indicator results of each model
表 1. 各模型的指标结果

模型/评价指标	Accuracy	Precision	Recall	IoU
PSPNet	96.88	91.16	91.18	83.77
DeepLabv3+	97.48	92.91	92.83	86.69
U-Net	97.76	94.11	93.13	88.00
SE-UNet	97.89	94.88	93.11	88.66
CBAM-UNet	97.97	94.91	93.50	89.04
本文模型	98.21	95.28	94.57	90.34

图 5 给出了几个视觉比较结果的例子，其中有 3 幅高分辨率遥感图像，包括不同形状、大小、分布模式和纹理特征的各种建筑。相比之下，PSPNet 的建筑物分割效果最差，一些建筑物没有被检测到，并且建筑物的边缘模糊不清；DeepLabv3+对建筑物的提取已经能取得不错的结果，但建筑物的边缘仍不准确；U-Net 不但能准确的提取建筑物的位置信息，而且对建筑物的边缘也有较好的分割效果。相较于 U-Net，本文提出的模型在建筑提取精度和视觉效果上都有更好的表现，尤其是在大型建筑内部的分割效果上，如图 5 中红色方块所标记的区域。



(a) 原始图像 (b) 标签 (c) PSPNet (d) DeepLabv3+ (e) U-Net (f) 本文模型

Figure 5. Comparison of building segmentation effects of different models
图 5. 不同模型的建筑物分割效果对比

4. 结论

本文提出了一种基于编码器–解码器结构并与 CSAM 注意力模块相结合的新型网络，用于从高分辨率遥感图像中精确提取建筑物。该网络通过加入注意力机制，抑制背景环境中的干扰因素，聚焦建筑信息，并增强多尺度、多层次的特征提取能力，提高了建筑物提取的精度。实验结果表明，该方法在 WHU 建筑物数据集上取得了较好的数值指标和视觉效果，像素精度、精准率、召回率和交并比分别达到了 98.21%、95.28%、94.57%和 90.34%。

参考文献

[1] 王俊, 秦其明, 叶昕, 等. 高分辨率光学遥感图像建筑物提取研究进展[J]. 遥感技术与应用, 2016, 31(4): 653-662+701.

[2] Al-Amri, S.S. and Kalyankar, N.V. (2010) Image Segmentation by Using Threshold Techniques. *Computer Vision and Pattern Recognition*, **2**.

[3] Liow, Y.T. and Pavlidis, T. (1990) Use of Shadows for Extracting Buildings in Aerial Images. *Computer Vision, Graphics, and Image Processing*, **49**, 242-277. https://doi.org/10.1016/0734-189X(90)90139-M

[4] Avudaiammal, R., Elaveni, P., Selvan, S., *et al*. (2020) Extraction of Buildings in Urban Area for Surface Area Assessment from Satellite Imagery Based on Morphological Building Index Using SVM Classifier. *Journal of the Indian Society of Remote Sensing*, **48**, 1325-1344. https://doi.org/10.1007/s12524-020-01161-0

[5] Long, J., Shelhamer, E. and Darrell, T. (2015) Fully Convolutional Networks for Semantic Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 7-12 June 2015, 3431-3440. https://doi.org/10.1109/CVPR.2015.7298965

[6] Badrinarayanan, V., Kendall, A. and Cipolla, R. (2017) Segnet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 2481-2495. https://doi.org/10.1109/TPAMI.2016.2644615

[7] Ronneberger, O., Fischer, P. and Brox, T. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention—MICCAI* 2015: 18*th International Conference*, Munich, 5-9 October 2015, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[8] Zhao, H., Shi, J., Qi, X., *et al*. (2017) Pyramid Scene Parsing Network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 2881-2890. https://doi.org/10.1109/CVPR.2017.660

[9] Chen, L.C., Papandreou, G., Kokkinos, I., *et al*. (2014) Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *Computer Vision and Pattern Recognition*.

[10] Chen, L.C., Papandreou, G., Kokkinos, I., *et al*. (2017) Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **40**, 834-848. https://doi.org/10.1109/TPAMI.2017.2699184

[11] Chen, L.C., Papandreou, G., Schroff, F., *et al*. (2017) Rethinking Atrous Convolution for Semantic Image Segmentation. *European Conference on Computer Vision*, 833-851.

[12] Chen, L.C., Zhu, Y., Papandreou, G., *et al*. (2018) Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *Proceedings of the European Conference on Computer Vision* (*ECCV*), 801-818. https://doi.org/10.1007/978-3-030-01234-2_49

[13] 李传林, 黄风华, 胡威, 等. 基于 Res__AttentionUnet 的高分辨率遥感影像建筑物提取方法[J]. 地球信息科学学报, 2021, 23(12): 2232-2243.

[14] Qiu, W., Gu, L., Gao, F., *et al*. (2023) Building Extraction from Very High-Resolution Remote Sensing Images Using Refine-UNet. *IEEE Geoscience and Remote Sensing Letters*, **20**, 1-5. https://doi.org/10.1109/LGRS.2023.3243609

[15] Hu, J., Shen, L. and Sun, G. (2018) Squeeze-and-Excitation Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake, 7132-7141. https://doi.org/10.1109/CVPR.2018.00745

[16] Woo, S., Park, J., Lee, J.Y., *et al*. (2018) CBAM: Convolutional Block Attention Module. *Proceedings of the European Conference on Computer Vision* (*ECCV*), 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[17] Hou, Q., Zhou, D. and Feng, J. (2021) Coordinate Attention for Efficient Mobile Network Design. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, 13713-13722. https://doi.org/10.1109/CVPR46437.2021.01350

[18] Ji, S., Wei, S. and Lu, M. (2018) Fully Convolutional Networks for Multisource Building Extraction from an Open Aerial and Satellite Imagery Data Set. *IEEE Transactions on Geoscience and Remote Sensing*, **57**, 574-586. https://doi.org/10.1109/TGRS.2018.2858817